



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

J. Michael Barbieri
03/02/2023



Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- In this project, we examined publicly available data from the web (Wikipedia) of SpaceX's rocket launching history to understand which features contributed to a successful Stage 1 rocket landing.
- Data were mined and assembled into a pandas dataframe using the Python coding language. Our target variable was the result of the landing: a variable called 'Class' with binary (0,1) values.
- Various data visualization techniques were employed to gain a better understanding of the independent variables, and the Class.
- Additionally, a web-app was constructed using some of the features available in the dataset, as well as an interactive dashboard.
- Finally, the data was prepared for various machine learning algorithms., including logistic regression, support vector machine, decision tree classifier, and k-nearest neighbors models.
- The data were standardized before being fed into the models, and categorical variables converted into dummy values.
- The results show that in terms of predicted future success rates of landing the Stage 1 rocket, the machine learning models all performed similarly; with an accuracy score of 83.33% across the board.
- This means that using the test data, there's a pretty good chance of determining what contributes to a successful landing; however, some limitations were present. Particularly, the lack of deviation in predictive quality across models, and presence of Type I error in predictions.

Introduction

- SpaceX has the potential to save over 100 million dollars with each rocket launch by recovering the "first stage" rocket; which has the capacity to come back down to earth and execute a propulsive landing. Since 2013, the success rate of Stage 1 rockets have increased; however, there are many variables involved in the process.
- Therefore, as scientists we would like to know what variables contribute to a successful Stage 1 landing. For example, the payload, orbital route, launch site, landing location (sea, land, etc.) all could play a role towards a successful landing, or not.
- Another consideration is the potential for competitors to replicate the success that SpaceX has been seeing. As private space exploration becomes more mainstream, the potential for up-and-coming companies to duplicate their accomplishments is increasing. So, it's important to use cutting edge technology (e.g., machine learning algorithms) to understand what's affecting the launching and landing sequences.

Section 1

Methodology

Methodology

Executive Summary

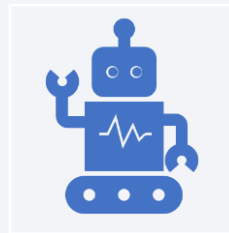
- Data collection methodology:
 - Data were collected from the SpaceX API, a publicly available data repository.
- Perform data wrangling
 - Data were checked for missing values and were supplied with the mean for that variable.
- Perform exploratory data analysis (EDA) using visualization and SQL
 - Visualizing the relationship between success rate and each orbit type; and the target Class with various other features (e.g., Flight Number, Orbit type, etc.)
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Data were collected from the SpaceX API, a publicly available data repository. Because the data-table included a lot of unnecessary information, only the pertinent data were passed into the dataframe.
- This process involved converting a JSON file-type into a Python pandas dataframe, filtering out the unnecessary information by mapping data into our list of features, and finally analyzing and dealing with null values in the dataframe.



SpaceX API



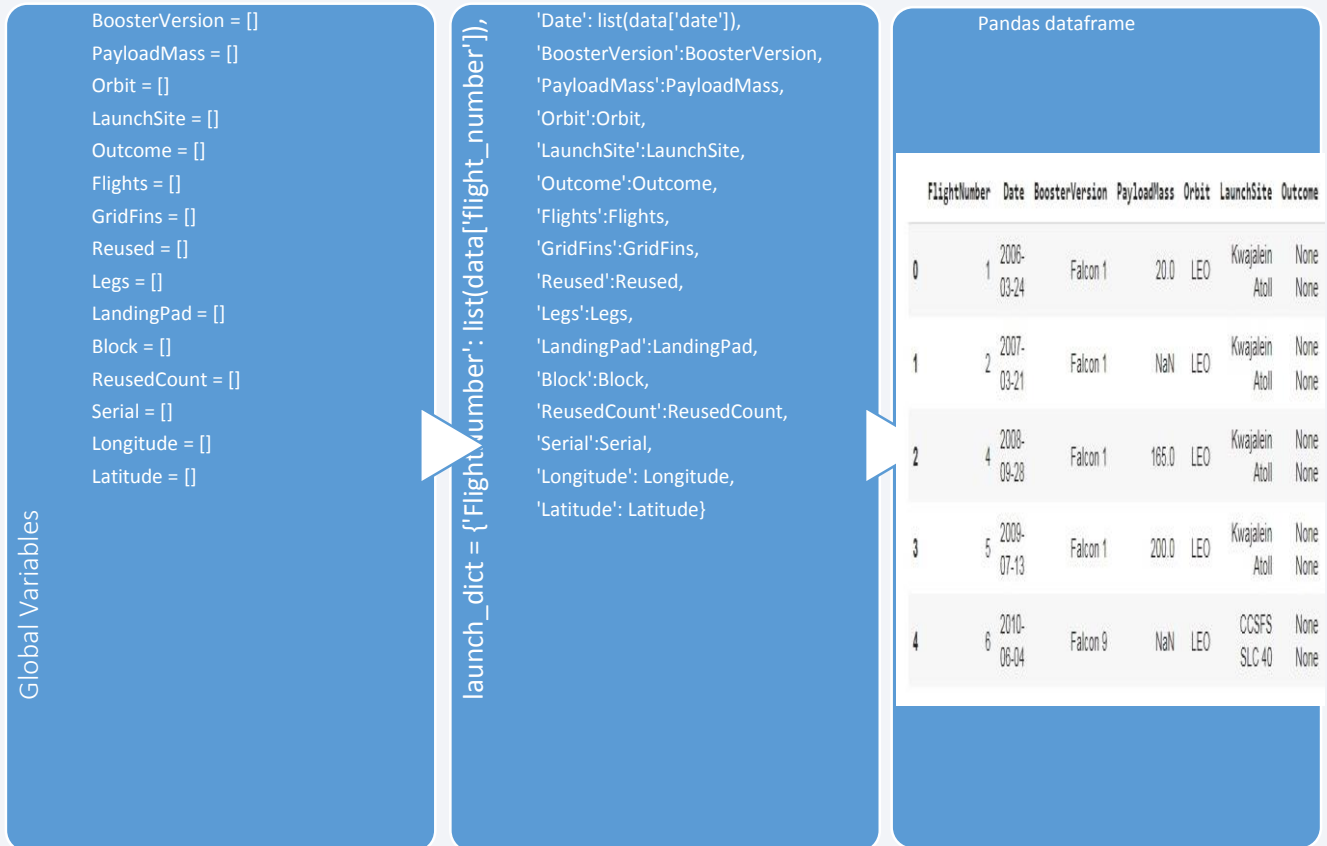
Data Scraping (Json
parsing & normalizing)



Dataframe Assembly
(data mapping onto lists)

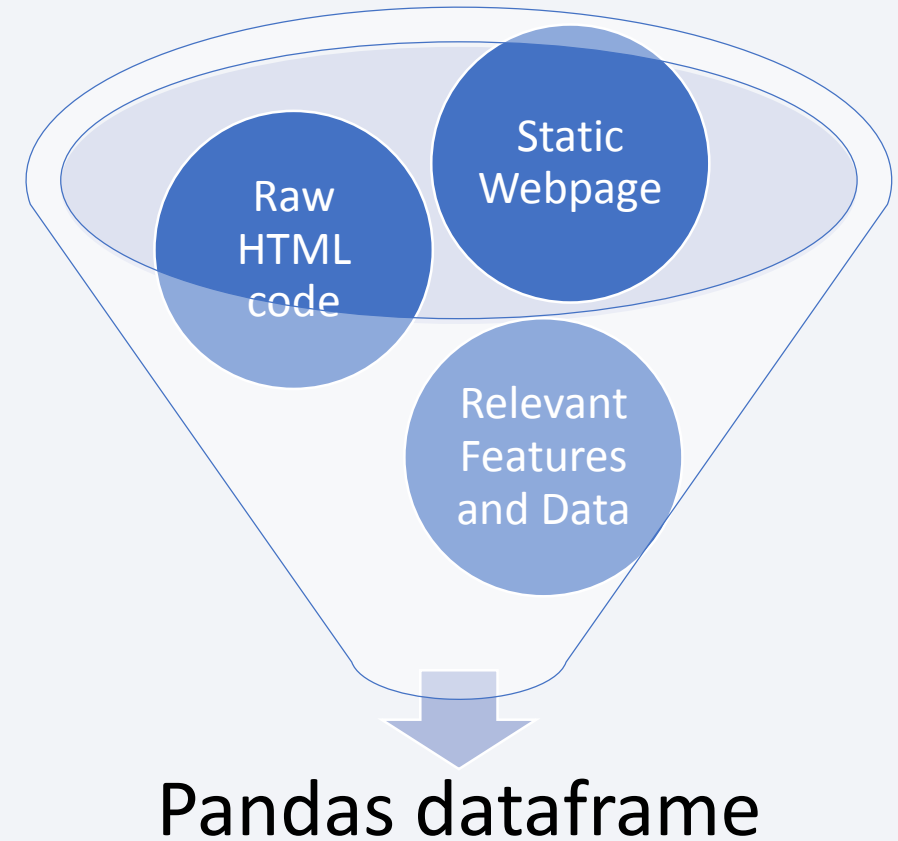
Data Wrangling – SpaceX API

- Once the dataframe was constructed, some further manipulation was required.
- The variable “Booster Version” was filtered to only include launches from the Falcon 9 rocket.
 - `data_falcon9 = df.loc[df['BoosterVersion'] != 'Falcon 1']`
- Missing values were discovered in the “Payload Mass” feature, which were replaced with that variable’s mean.
 - `data_falcon9['PayloadMass'].fillna(payload_mean, inplace = True)`
- Finally, the entire dataframe was cast to a csv file, and stored remotely.
- A copy of this lab can be found at the following link:
 - [Coursera/spacex api.ipynb at main · Barbj379/Coursera \(github.com\)](https://www.coursera.org/learn/spacex-api-ipython-at-main-Barbj379/Coursera/github.com)



Data Collection - Scraping

- To collect records of historical Falcon 9 launches, web-scraping processes were conducted using the BeautifulSoup API.
- These launch records are stored in [a HTML table at Wikipedia](#), and were retrieved from the attached static url above.
- The target table is the third table available at that webpage, which was targeted for this analysis.
- The raw HTML data had to be parsed through to discover the column names, which were appended to dictionary.
- This dictionary was then supplied with the appropriate data to fully flesh-out the dataframe.
- [Coursera/web scraping lab.ipynb at main · Barbj379/Coursera \(github.com\)](#)



EDA with Data Visualization

- Exploratory Data Analysis (EDA) was conducted using the matplotlib and seaborn libraries.
- The primary target variable is 'Class', a feature that was constructed using binary responses of launch outcomes: 1 being success, and 0 being unsuccessful rocket landing.
- So, the relationship between the type of Orbit and the success rate was examined with a barchart, as well as success rates' yearly trend over time.
- Additional analyses included scatterplots of the Payload Mass and Orbit type; Flight Number and Orbit type; Payload Mass and Launch Site, and Flight Number and Launch Site.
- [Coursera/data_viz_lab.ipynb at main · Barbj379/Coursera \(github.com\)](#)

EDA with SQL

- Further EDA was conducted using the following SQL queries:
 - Total payload mass carried by boosters launched by NASA.
 - Average payload mass by a specific booster (F9 v.1.1)
 - The date of the first successful landing outcome on a ground pad.
 - The names of boosters which have had successful landings on drone ships, and a payload mass between 4,000 and 6,000 kg.
 - The total number of successful and failure mission outcomes.
 - The names of the booster types that have carried the maximum payload.
 - The records that contain a failed landing outcome on a drone ship, including month name and launch site in 2015.
 - Ranking the count of successful landing outcomes between 04/06/2010 and 03/20/2017.
- [Coursera/sqk lite.ipynb at main · Barbj379/Coursera \(github.com\)](#)

Build an Interactive Map with Folium

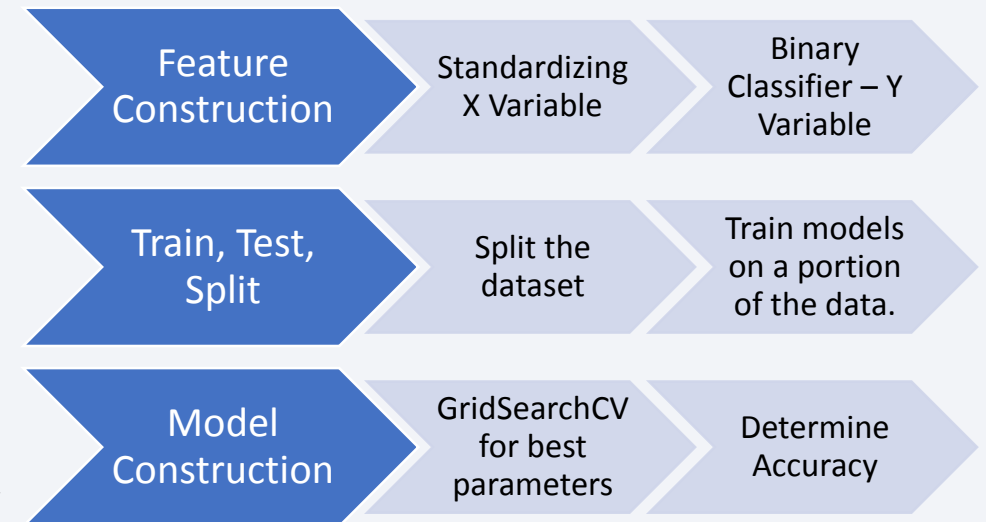
- An interactive map was constructed using Folium to understand some relevant information about the launch sites themselves.
- For example, these sites were plotted using their latitudes and longitudes to know what kinds of topographical features are associated with each site.
- Markers were placed at each site to delineate both successful and failed rocket launches; these were coded in red and green, and organized by cluster.
- Furthermore, the proximity of launch sites to various structures was learned:
 - Distance from the launch site to nearest railway, nearest highway, and nearest city, as well as calculated distance to the coastline were plotted using Polyline.
- [Coursera/map_folium_lab.ipynb at main · Barbj379/Coursera \(github.com\)](#)

Build a Dashboard with Plotly Dash

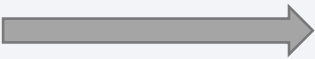
- An interactive dashboard was constructed using Plotly Dash to better understand some of the functional relationships between variables.
- For example, a dropdown menu with all launch sites was constructed to display success rates using a pie chart.
- These data can be further filtered by the payload range of each launch at the various sites.
- This provides an insight into the success rate of different rockets carrying different payloads at the various launch sites.
- [Coursera/working_spacex_dash_app.py at main · Barbj379/Coursera \(github.com\)](https://github.com/Barbj379/Coursera_working_spacex_dash_app.py)

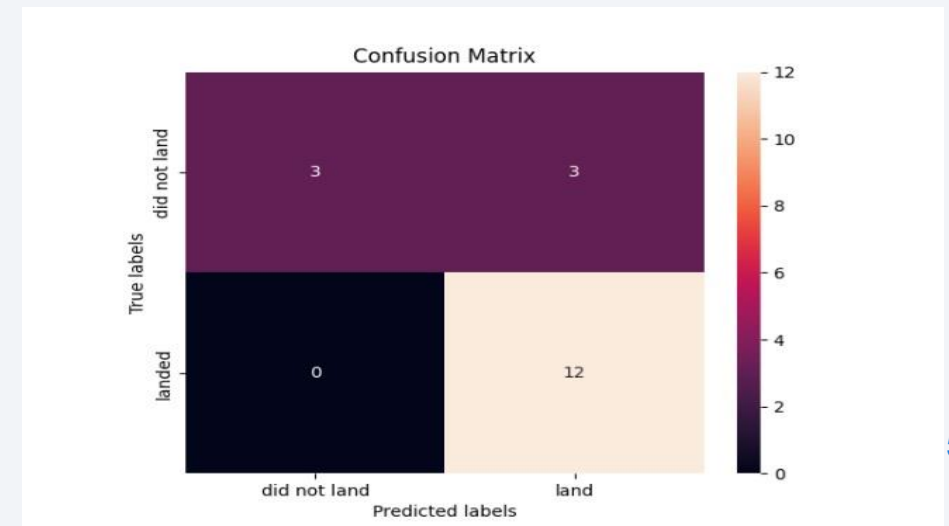
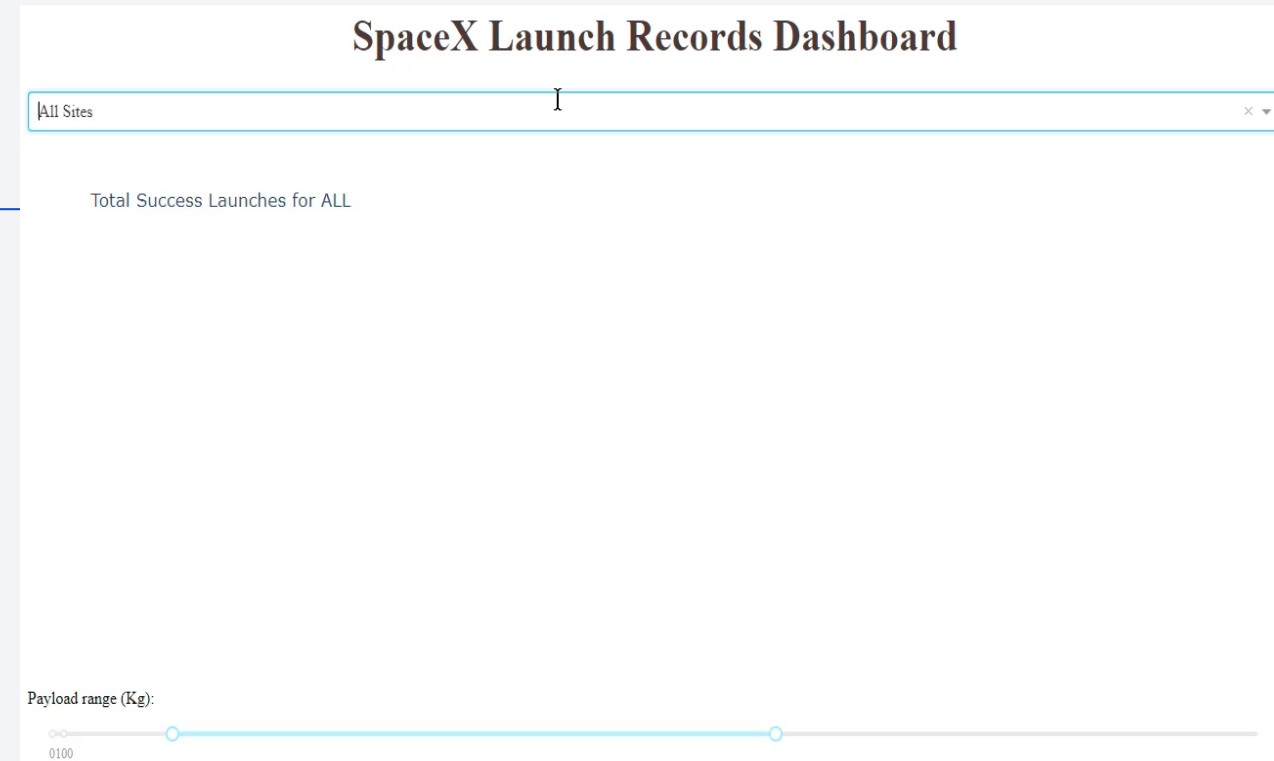
Predictive Analysis (Classification)

- To make predictive analyses, various machine learning models had to be trained, tested, evaluated, and improved to determine the best prediction accuracy for successful launches.
- X – the independent variables were standardized by transformation with `.StandardScaler()`
- Y – the target variable was encoded as a binary response: 1 for success, 0 for failure.
- Train, test, split function allowed each of the models to learn from a portion of the dataset.
- Then the training data were fed into each of the following algorithms:
 - Logistic Regression, Support Vector Machine, Decision Tree classifier, K-Nearest Neighbors
- The accuracy of each model was calculated by the `.score` method, and a confusion matrix was plotted for each model.
- The GridSearchCV method determined the best parameters to use for each of the models.
- [Coursera/predictive_analysis.ipynb at main · Barbj379/Coursera \(github.com\)](#)



Results

- EDA contributed meaningful insights that will be explored in following slides:
 - Different orbit types have greater success rates.
 - Different payloads can constrain orbit type.
 - Increasing trend of success rate since 2013.
 - Different launch sites host various payloads.
- All of the machine learning models happen to score an 83.33% accuracy for predicting successful launches.
- Each of the results presented the following confusion matrix: 



The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

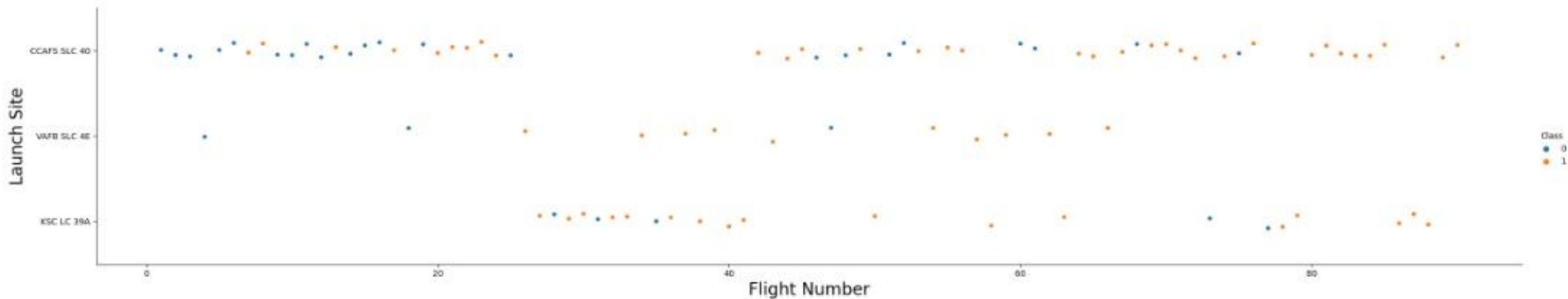
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

TASK 1: Visualize the relationship between Flight Number and Launch Site

```
### TASK 1: Visualize the relationship between Flight Number and Launch Site
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



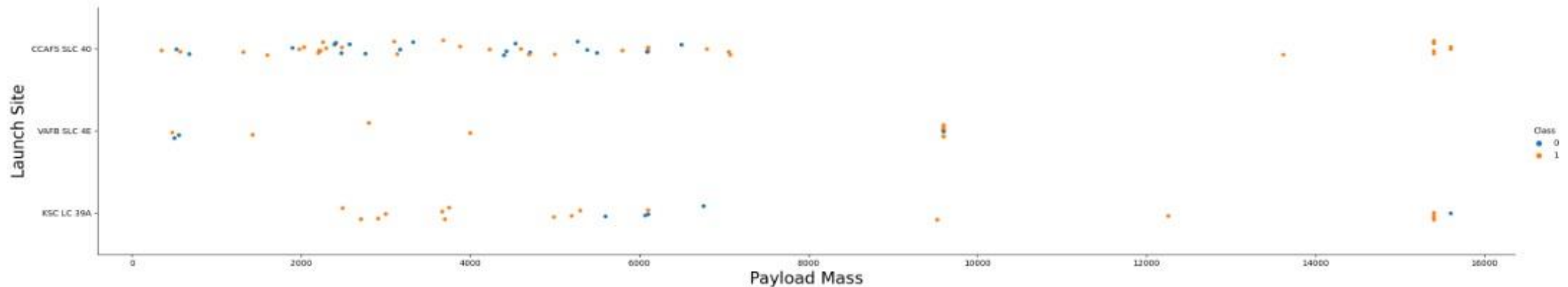
- Some launch sites have more launches than others.
- Site CCAFS SLC-40 seems to have more failures early on, and more success later on.
- Sites KSC LC-39A and VAFB SLC 4E have a success rate of 77%, (compared to 60% for CCAFS LC-40).

Payload vs. Launch Site

TASK 2: Visualize the relationship between Payload and Launch Site

Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

```
1: ### TASK 2: Visualize the relationship between Payload and Launch Site
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

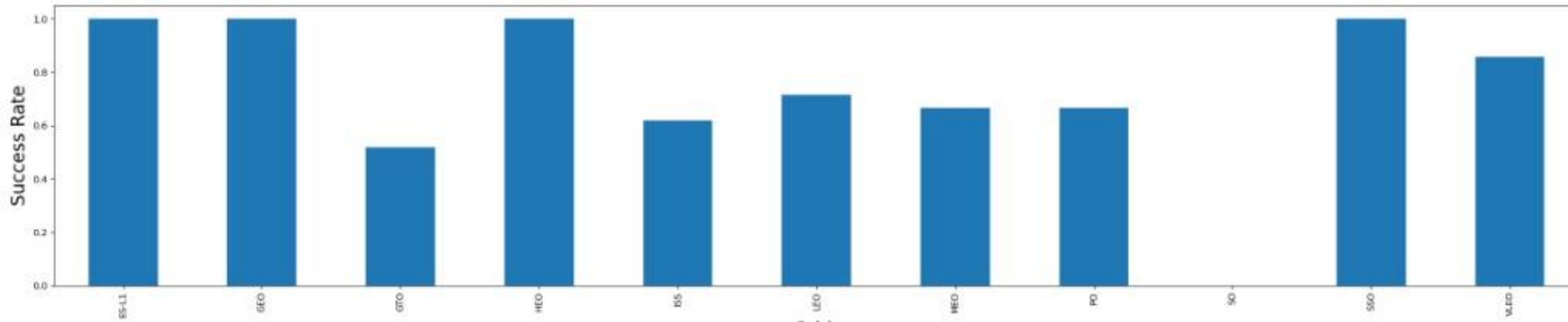


- There aren't any rockets weighing greater than 10,000kg launched from VAFB-SLC.
- Most of the lighter payloads (<10,00kg) were launched at CCAFS SLC-40.

Success Rate vs. Orbit Type

TASK 3: Visualize the relationship between success rate of each orbit type

```
#orbit = df.groupby(['Orbit'])['Class'].mean()
df.groupby("Orbit").mean()['Class'].plot(kind='bar')
plt.xlabel("Orbit",fontsize=20)
plt.ylabel("Success Rate",fontsize=20)
plt.show()
```

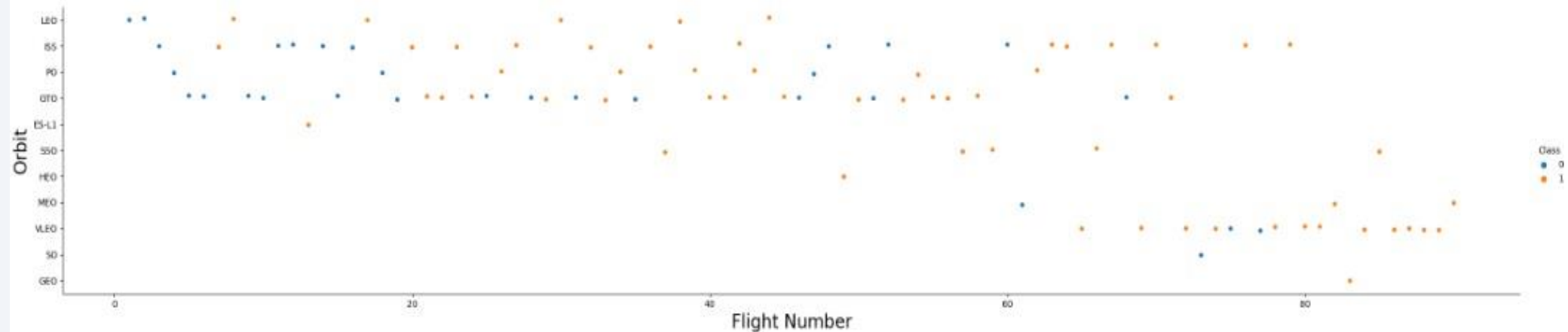


- Different orbit types seem to have better success rates.
- The following orbits have a perfect record at the time of data collection:
 - ES-11, GTO, HEO, SSO

Flight Number vs. Orbit Type

TASK 4: Visualize the relationship between FlightNumber and Orbit type

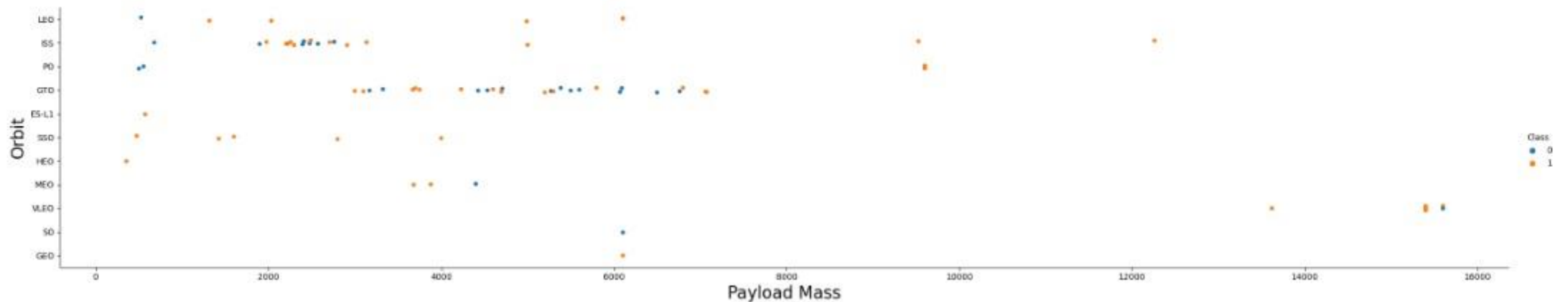
```
### TASK 4: Visualize the relationship between FlightNumber and Orbit type
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Orbit", fontsize=20)
plt.show()
```



- In regard to the LEO orbit, the Success frequency appears related to the number of flights; however, this isn't apparent between flights in the GTO orbit.

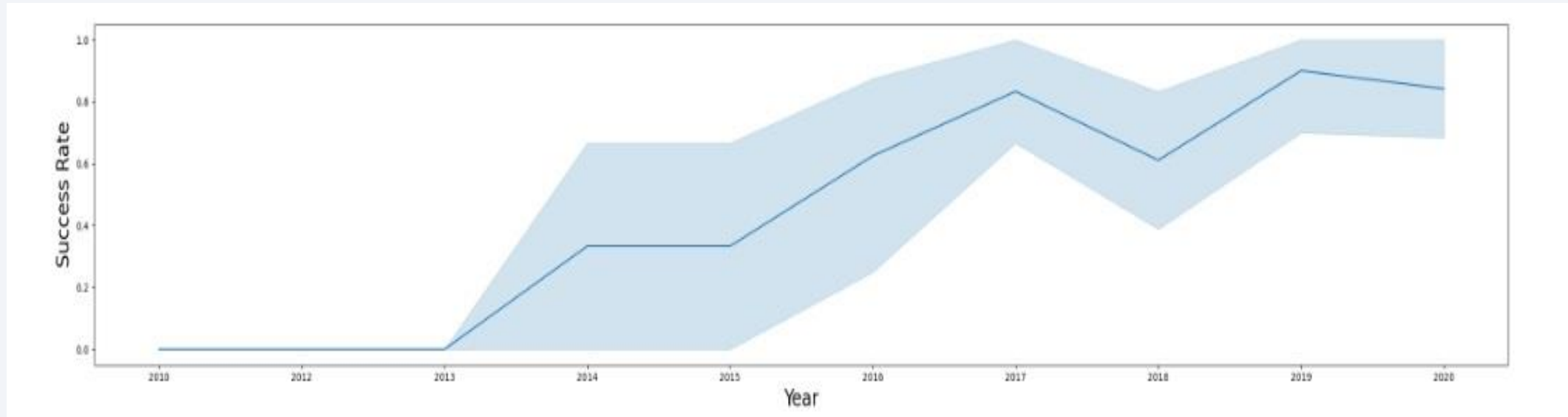
Payload vs. Orbit Type

```
58]: sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Payload Mass",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



- With heavy payloads, the Polar, LEO and ISS orbits experience more frequent successful landings.
- For GTO there are a high frequency of flights, including both successful and failed landings.

Launch Success Yearly Trend



- The success rate for SpaceX's Stage1 booster rockets has increased steadily from 2013 to 2020.

EDA using SQL --- All Launch Site Names

- Using SQL, the launch site names include:
 - CCAFS LC-40
 - VAFB SLC-4E
 - KSC LC-39A
 - CCAFS SLC-40

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Here are five records from the dataset that include a launch site beginning with the characters 'CCA.' All other information in the record is retained too.

Total Payload Mass

- The total payload mass carried by rockets from NASA was found to be 99980.
- This is a lot of weight!

```
sum("PAYLOAD_MASS_KG_")
```

```
99980
```

Average Payload Mass by F9 v1.1

```
avg("PAYLOAD_MASS_KG_")
```

```
2534.6666666666665
```

- The average payload mass carried by booster version F9 v1.1 was found to be 2534.66kg. This is on the lighter side of the range.

First Successful Ground Landing Date

- Using the Min() function in SQL, the first successful ground landing date was discovered to be on January 5th, 2017.

MIN("Date")

01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters that experienced a successful landing on a drone ship, carrying a payload between 4,000-6,000kg were queried.
- The results are presented here.

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes were counted using SQL.
- The results are presented here.

<code>count("Mission_Outcome")</code>
101

Boosters Carried Maximum Payload

- The unique names of the boosters which have carried the maximum payload mass was queried using SQL.
- The results are presented here.

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- In 2015, two boosters failed to land successfully on drone ships. The booster version, and pertinent data, are presented here.

month	Booster_Version	Landing_Outcome
01	F9 v1.1 B1012	Failure (drone ship)
04	F9 v1.1 B1015	Failure (drone ship)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- The count of successful landing outcomes between 06/04/2010 and 03/20/2017, in descending order, are presented here.

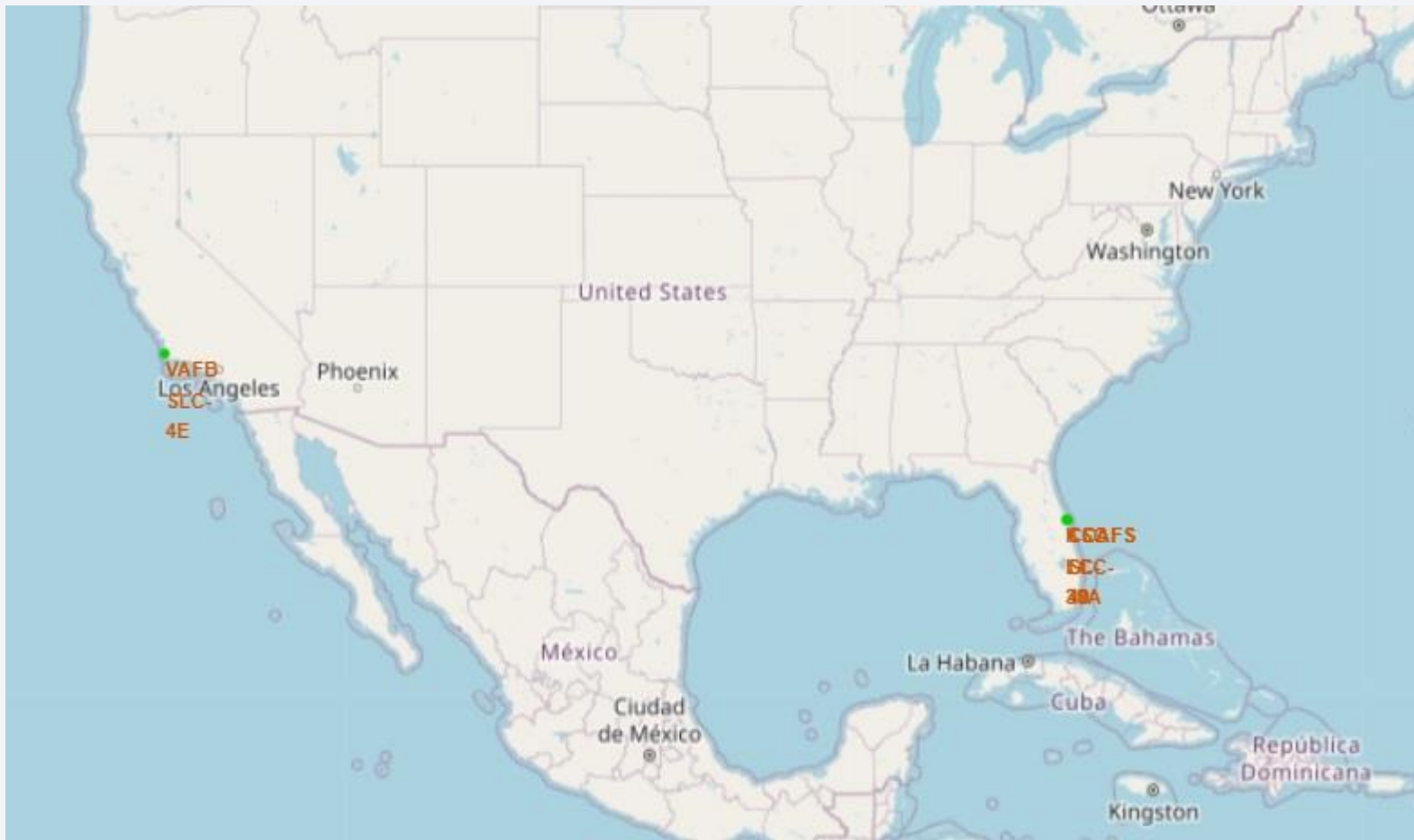
count("Mission_Outcome")	
	20
	10
	8
	6
	4
	3
	3
	2
	1

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

Launch Sites Proximities Analysis

Interactive Geographical Map



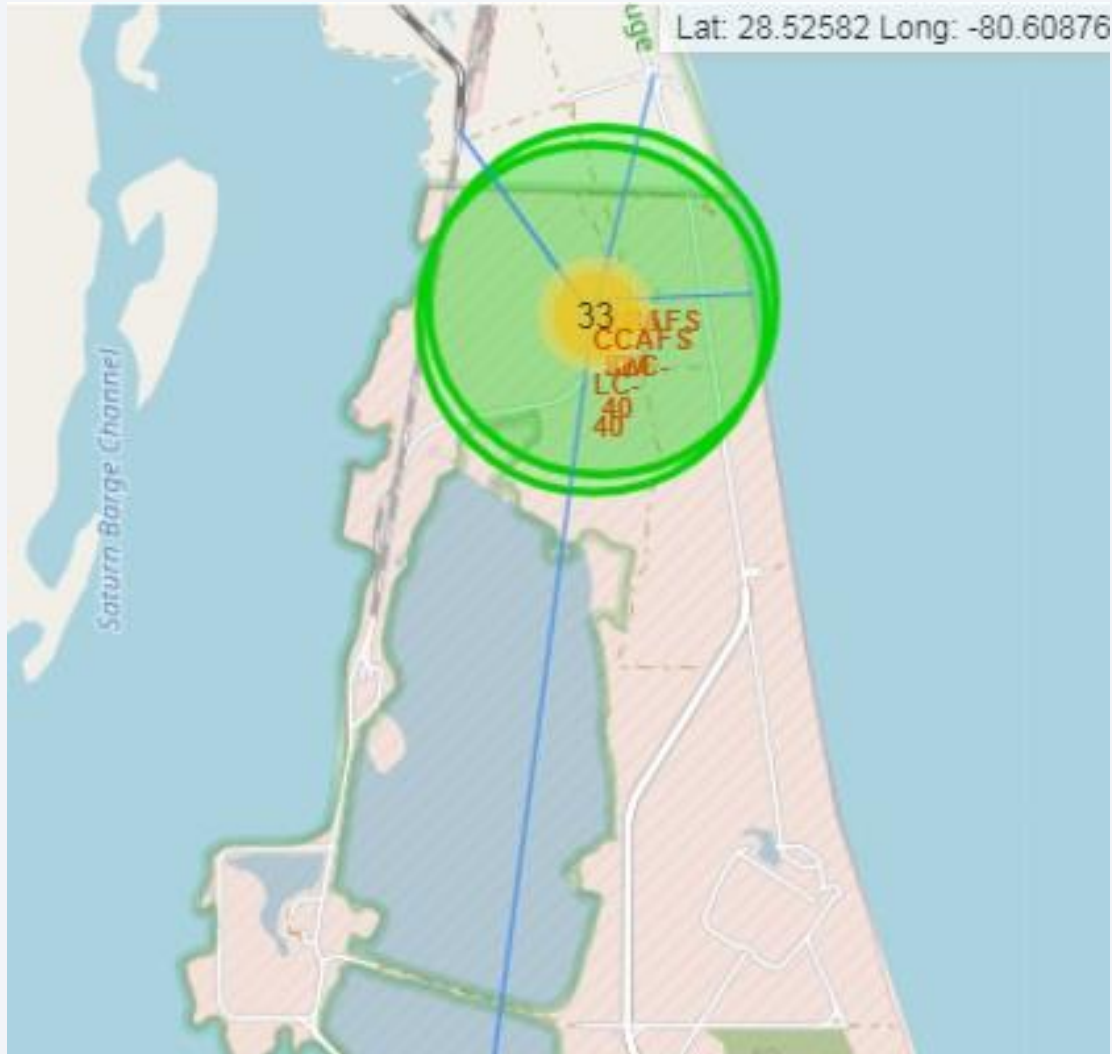
- Each of the SpaceX launch sites were plotted on a geographical map using the folium library.
- They're represented by a green circle.

Interactive Geographical Map, Cont'd.



- Here, additional information is provided to the map, including the frequency of successful launch outcomes per site.
- Each circle can be zoomed-in upon for more detail.

Interactive Geographical Map, Cont'd



- Here, more nuanced information is given from drawing lines using Polyline function to nearby structures.
- For example, the distance to the coastline, a railway, a highway, and major city are drawn on this map.
- Also, distance to the coastline was calculated.

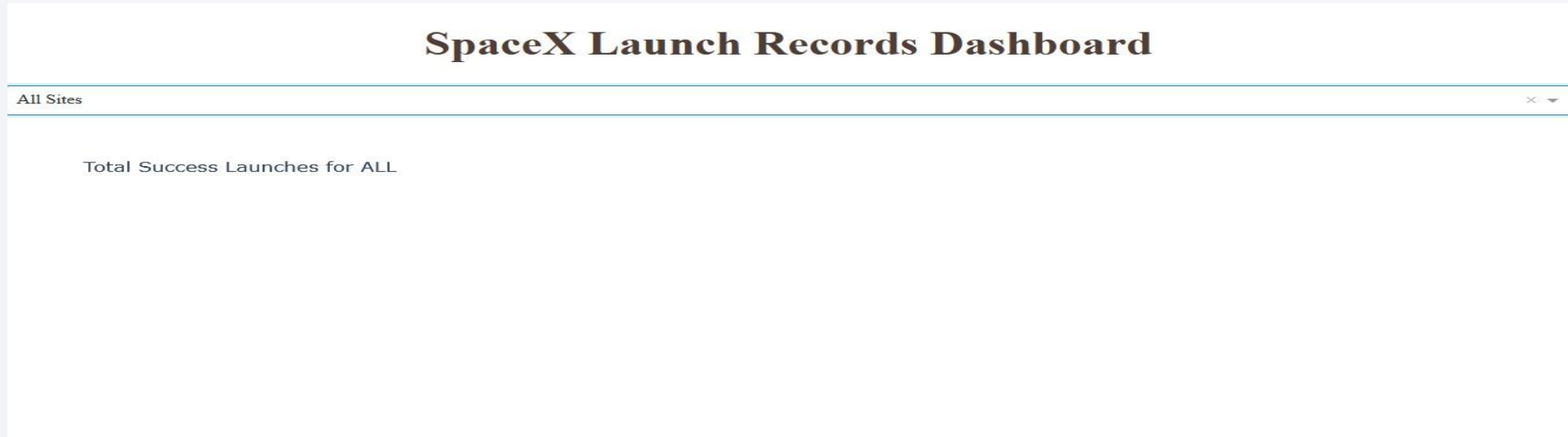
0.8512477740804306



Section 4

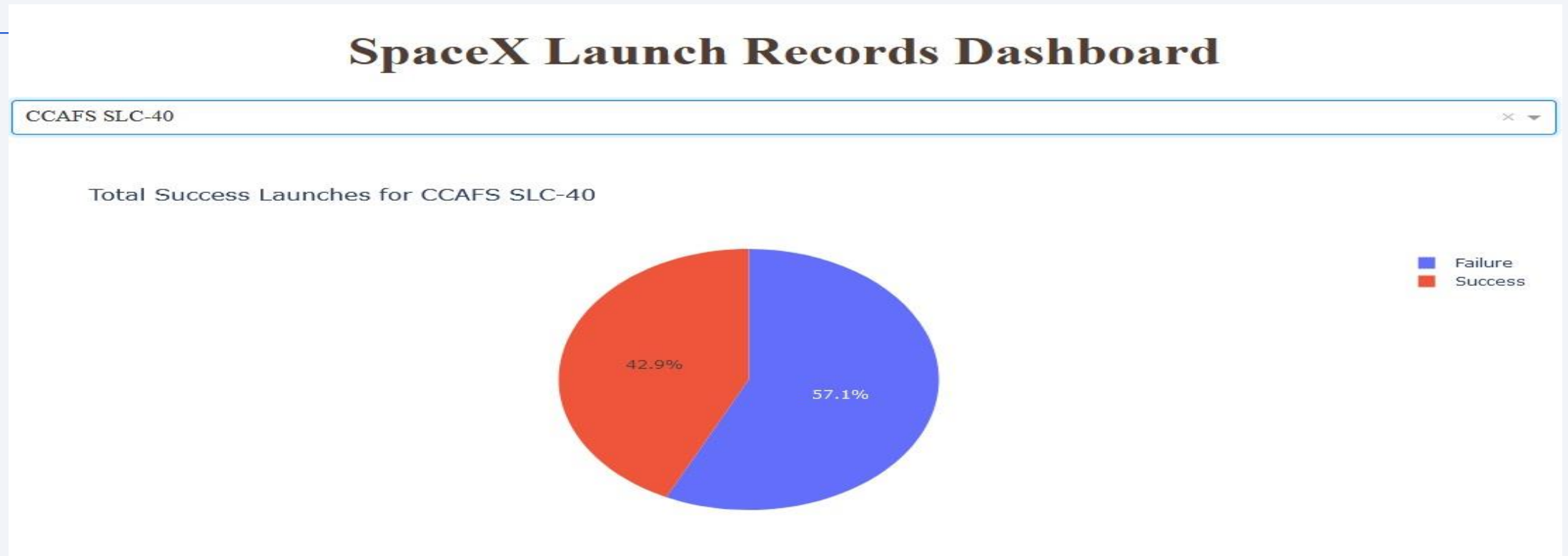
Build a Dashboard with Plotly Dash

SpaceX Dashboard



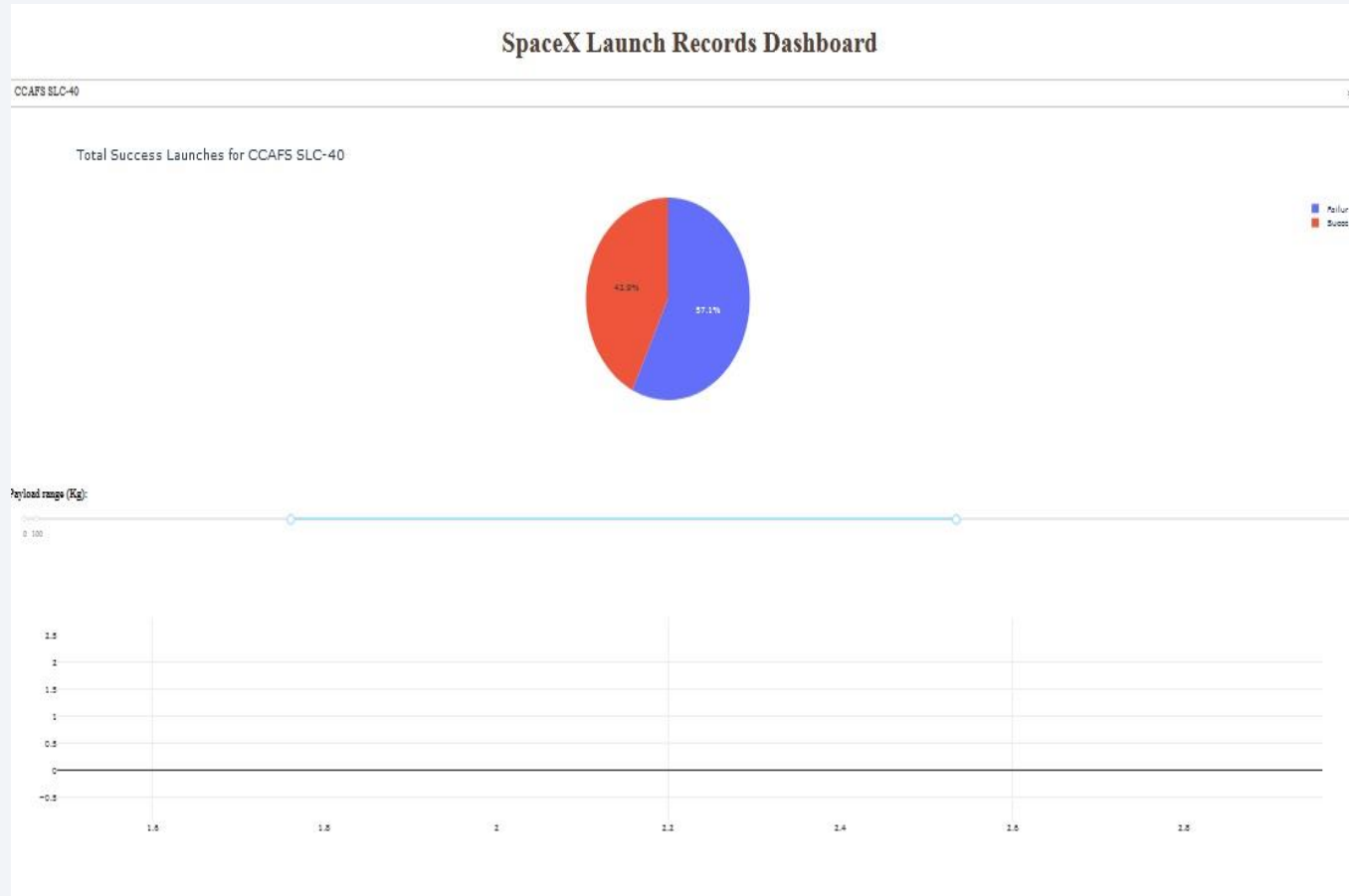
- The launch success for all sites was selected in the web-app created from Plotly, [here](#).

SpaceX Dashboard, Cont'd



- The site with the greatest percentage of successful launches is CCAFS SLC-40, with a 42.9% success rate.

SpaceX Dashboard, Cont'd.



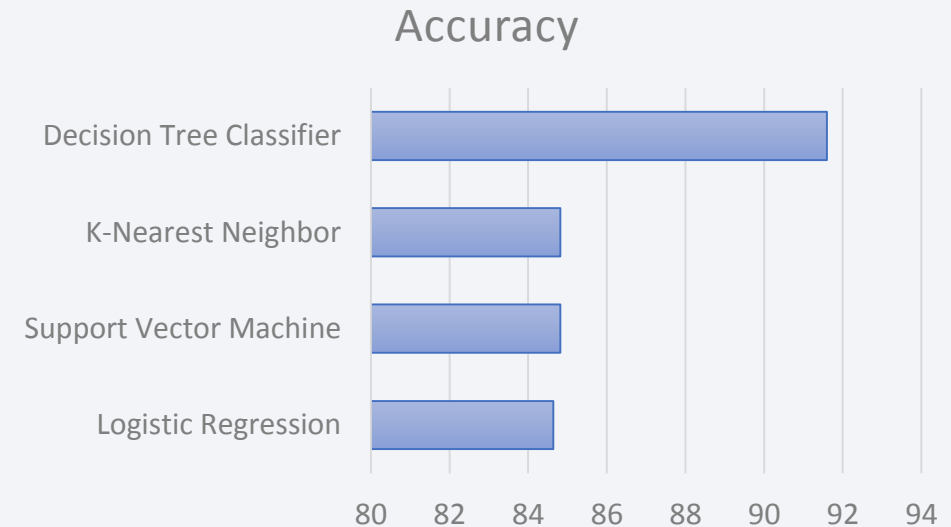
- The payload range and launch site with highest success rate are presented here.

Section 5

Predictive Analysis (Classification)

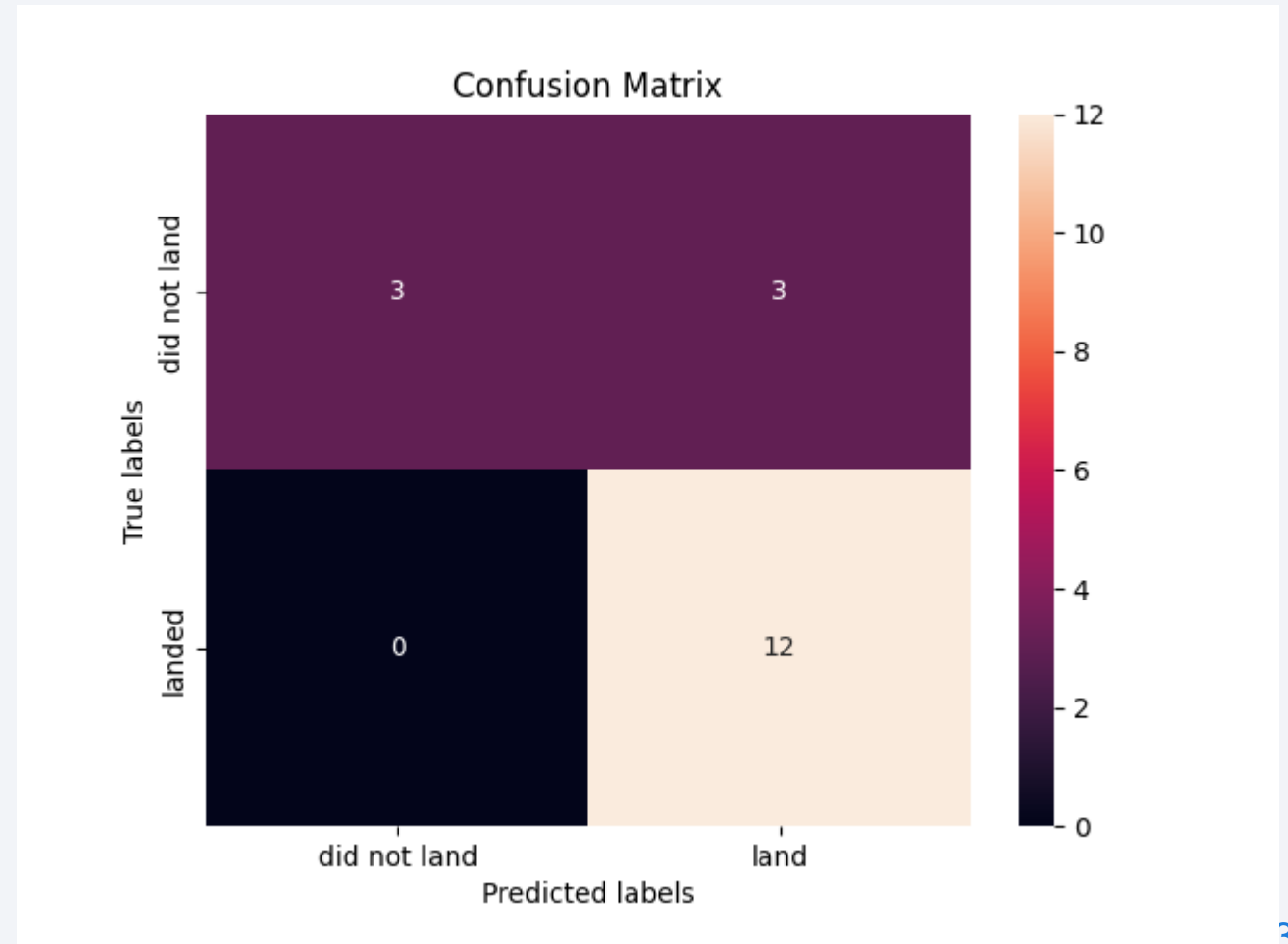
Classification Accuracy

- In terms of accuracy of the various models on the test data, each of the four models scored the same: 83.33% accuracy.
- But after each model was hypertuned to fit the best parameters, the Decision Tree Classifier presented the highest accuracy on the validation data: 91.6% accuracy.



Confusion Matrix

- Here is the decision tree classifier model's confusion matrix. The model scored very well on prediction true positives (12/12 in lower right quadrant).
- But, like the other models, it had quite a few false positives (3/3 in the upper right quadrant). So, it thought the rocket would land safely when it didn't in reality.



Conclusions

- In conclusion, the Space Race has intensified significantly in recent years with technology becoming more abundant and affordable; and space travel becoming more mainstream.
- Although a very costly pursuit still, SpaceX has engineered a Stage1 rocket that can land safely back on earth after it's orbit. This reduces the cost of each launch considerably.
- To better understand what factors are associated with a successful landing, this project used publically available data to assemble a dataframe that was fed into various machine learning algorithms.
- The project also employed data visualization and SQL querying techniques to get a better understanding of which factors influence a safe and successful landing.
- It was shown that overall, the ML models could predict if a Stage 1 rocket would land safely back on earth with an 83.33% accuracy. This information can be used to further explore the many different variables that are involved with space travel.

Appendix

- [Github repository with all the relevant labwork.](#)
- Python libraries used:
 - %load_ext sql
 - Pandas
 - Matplotlib
 - Seaborn
 - Folium
 - Sklearn – StandardScaler, Train_test_split, GridSearchCV, LogisticRegression, SVC, DecisionTreeClassifier, KNeighborsClassifier
 - BeautifulSoup
 - Dash – html, dcc, Input, Output, Plotly.express

Thank you!

